

Research Data Management

Proč a jak?



**FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE**

doc. Ing. Robert Pergl, Ph.D.
robert.pergl@fit.cvut.cz
robert.pergl@ds-wizard.org



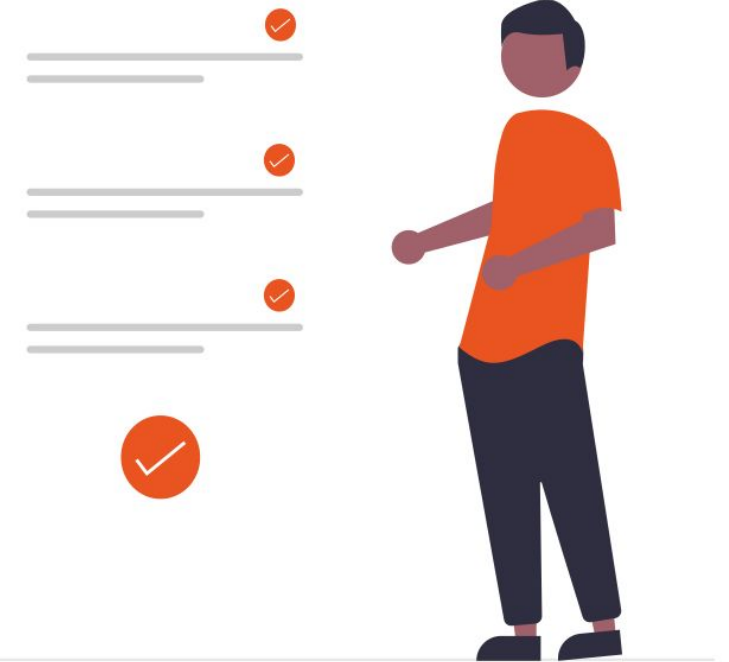
doc. Ing. Robert Pergl, Ph.D.

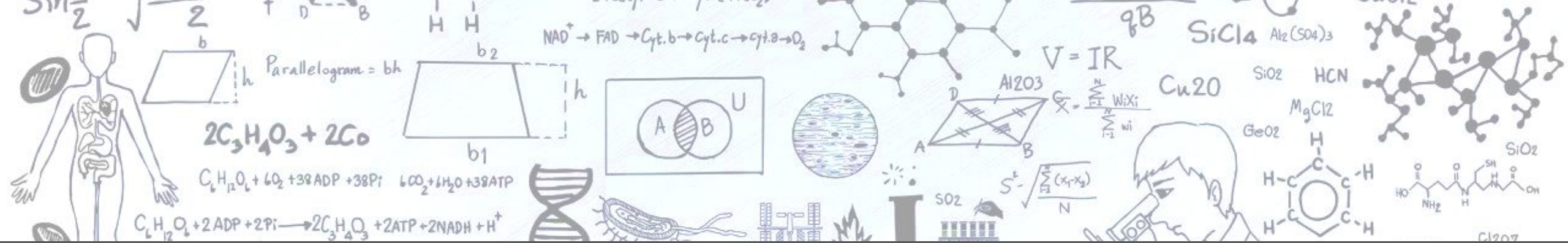
- vedoucí Centra pro konceptuální modelování a implementace na FIT ČVUT ([CCMi](#))
- zástupce ve Výboru ELIXIR CZ za ČVUT
- projektový koordinátor *Data Stewardship Wizard*

robert.pergl@fit.cvut.cz

robert.pergl@ds-wizard.org

- Research Data Management
- FAIR
- Data Management Plan
- Data Stewardship Wizard

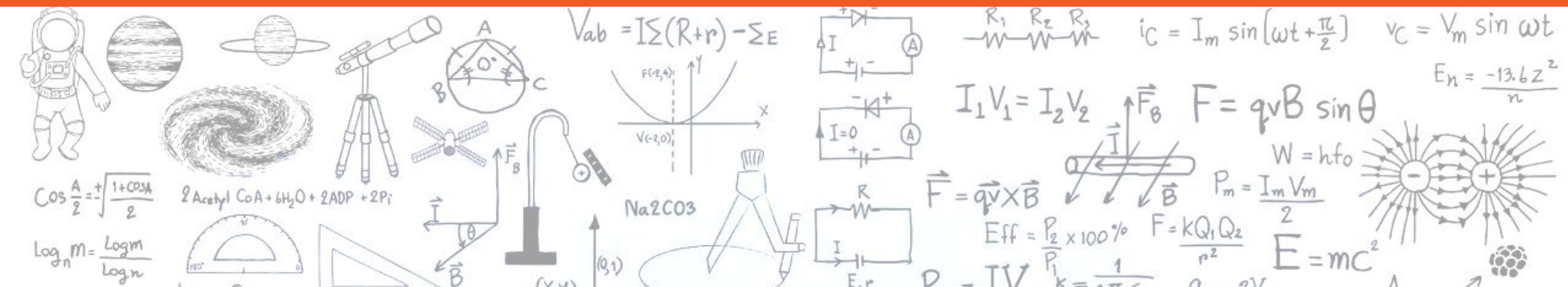




DSW

DATA STEWARDSHIP WIZARD

Research Data Management



Trocha motivační historie...

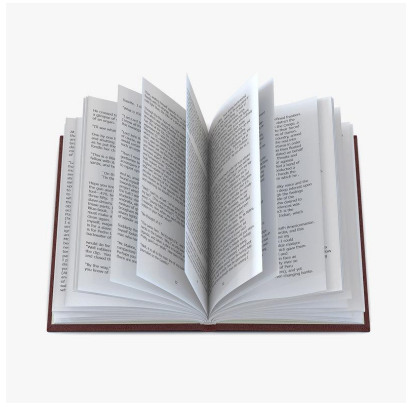
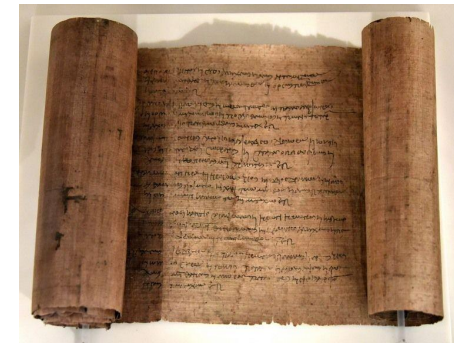
Práce s daty (a zejména jejich sdílení!) stojí za úspěchem Homo Sapiens.



[Why did humans prevail?](#)

Trocha motivační historie...

Historie lidstva je i historií sběru, reprezentace, interpretace, zpracování, ukládání, využívání a sdílení dat.



Trocha motivační historie...



Historie lidstva je i historií sběru, reprezentace, interpretace, zpracování, ukládání, využívání a sdílení dat.



Microsoft Excel spreadsheet showing a personnel list for Childs Play, Inc.

Emp ID	Last Name	First Name	Position	Department	Division	Salary	Start Date	Birth Date
1011	Gorton	Hazel	Accounting Assist.	Accounting	Toys	\$29,585	2/3/1986	11/21/1964
1012	Preston	Liza	Mechanical Engineer	Engineering	Games	\$43,394	1/26/1986	12/2/1964
1041	Tercan	Robert	Group Admin. Assist.	R and D	Games	\$28,044	4/16/1992	1/25/1965
1054	Smith	Howard	Design Assist.	Art	Toys	\$25,176	4/16/1991	8/9/1967
1055	Albert	Maxine	Group Admin. Assist.	Marketing	Toys	\$31,678	4/8/1991	8/20/1967
1056	Gonzales	Joe	Unit Mgr.	Admin.	Toys	\$116,511	10/25/1979	8/24/1937
1067	Scote	Gail	Design Specialist	Art	Teaching Aids	\$36,940	9/20/1987	9/30/1961
1068	Mann	Alyssa	Mechanical Engineer	Engineering	Games	\$47,883	9/12/1987	10/11/1961
1076	Kane	Sheryl	Design Assist.	Art	Games	\$23,239	8/7/1992	8/28/1969
1076	McCormick	Molly	Lead Engineer	Engineering	Toys	\$105,753	7/30/1979	9/8/1940
1078	Hapabuch	Kendrick	Admin. Assist.	Marketing	Games	\$29,983	4/1/1986	11/21/1962
1079	Pirce	Ellen	Admin. Assist.	Admin.	Games	\$29,983	3/24/1986	12/6/1962
1080	Foss	Felix	Research Scientist	R and D	Games	\$64,738	10/29/1988	12/6/1952
1152	Henders	Mark	Accounting Assist.	Accounting	Games	\$26,646	1/21/1990	10/23/1965

The screenshot shows the PDB website interface with search results for 'HAEMOGLOBIN'. It includes filters for Organism (Homo sapiens, E. coli, Bacteria), X-Ray Resolution, Release Date, and Polymer Type. The search results show 89 structure hits.

Trocha motivační historie...

A též rozvojem systémů správy dat (data management)



- Výrazně rostoucí objem vědeckých dat představuje technické, ale i organizační výzvy.
- Data přináší společnosti výraznou hodnotu – jsou podkladem pro analýzy, rozhodování (viz např. COVID).
- Data se stávají klíčovým motorem pokroku (nejen) ve vědě.
- Správa dat je tak jedním z hlavních aktuálních témat v administrativě vědeckých projektů.

Životní cyklus dat (Data Life Cycle)



Data Life Cycle: Plan



- Plánování jakým způsobem se bude během projektu nakládat s daty
- Výstupem by měl být **Data Management Plan** (více dále)

Data Life Cycle: Collect



- Sběr nových dat (metody se liší podle výzkumné oblasti)
- Použití existujících dat (např. z předchozích projektů)
- Důraz na kvalitu použitých dat
- Zaznamenání původu (*provenance*) dat - kdo, pomocí čeho (nástroje), podmínky experimentu, atd.

Data Life Cycle: Process



- Převod dat z uloženého formátu do formátu vhodného pro analýzu
- Vyřazení špatných dat nebo dat z nízkou kvalitou
- Pseudonimizace/anonymizace citlivých dat

Data Life Cycle: Analyse



- Zkoumání nasbíraných dat
- Hlavní část výzkumu – získávání nových znalostí
- Workflow použité v analýze by mělo být reprodukovatelné
- Analýza velkých dat může vyžadovat velký výpočetní výkon
- Potřeba specializovaných oborových softwarových nástrojů (viz např. bio.tools)
- K dispozici jsou [velké výzkumné infrastruktury](#), např. [ELIXIR](#), [CLARIN](#)

Data Life Cycle: Preserve



- Zajištění dlouhodobého uchování dat po ukončení projektu:
 - Možnost ověření výsledků projektu i po letech
 - Využití dat v budoucnu pro jiné účely (výuka, další výzkum)

Data Life Cycle: Share



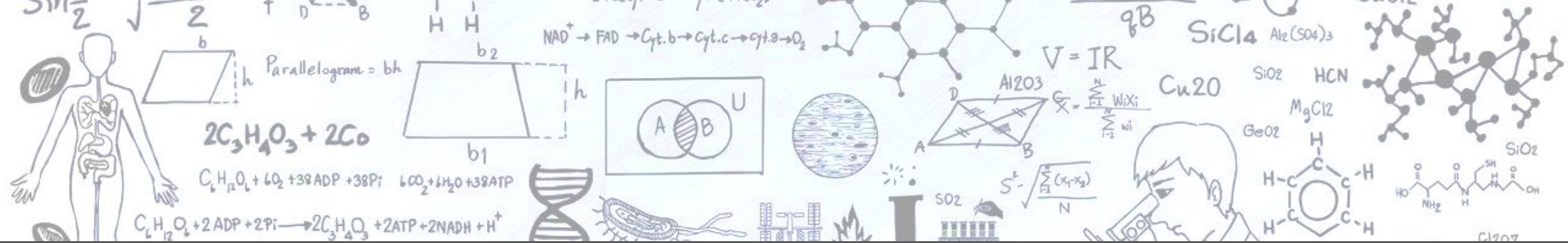
- Sdílení dat s ostatními (např. na dalším výzkumném projektu)
- Sdílení neznamená, že data musí být vždy veřejně dostupná (open), mohou být sdílená pouze za určitých (omezujících) podmínek a přístup může být i placený
- Zvážení všech etických, právních, licenčních a jiných omezení
- Princip "As open as possible, as closed as necessary" (Evropská komise)

Data Life Cycle: Reuse



- Použití dat pro jiný účel, než pro který byla nasbírána, např.:
 - Jako referenční data pro jiný výzkum
 - Ověření výsledků původního výzkumu
 - Propojování výsledků více studií do meta-studií

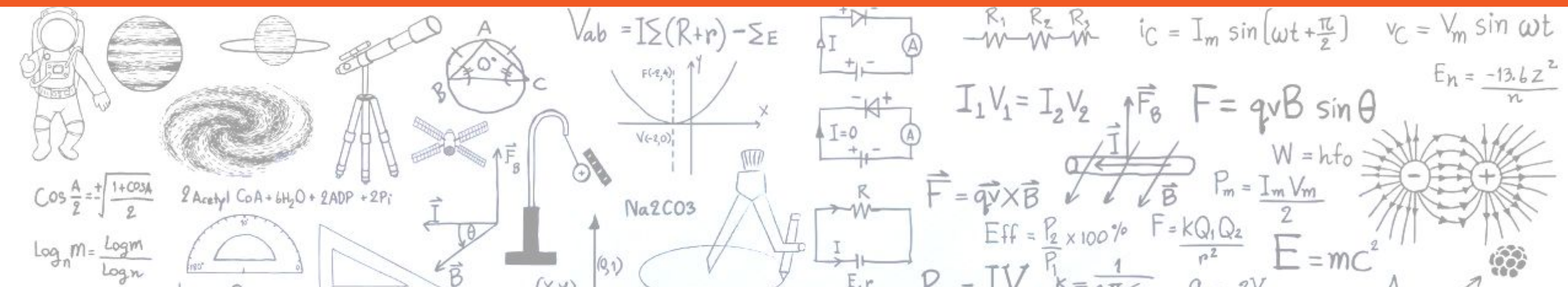
- **Výzkumník (Researcher)** = pracuje s daty během výzkumu, píše DMP pro svůj projekt
 - Výzkumníci, PhD studenti, žadatelé o granty, projektoví manažeři, atd.
- **Správce dat (Data Steward)** = stará se o práci s daty na různých úrovních
 - **Policy** = spolupráce s vedením a grantovými agenturami; dohled na správný postup, etické a právní záležitosti
 - **Research** = spolupráce s vědci; kontrola a pomoc s tvorbou DMP
 - **Infrastructure** = spolupráce s IT; IT řešení pro RDM, infrastruktura, atd.
 - Někdy ještě [jemnější dělení na různé role](#) (data custodian, data curator, apod)

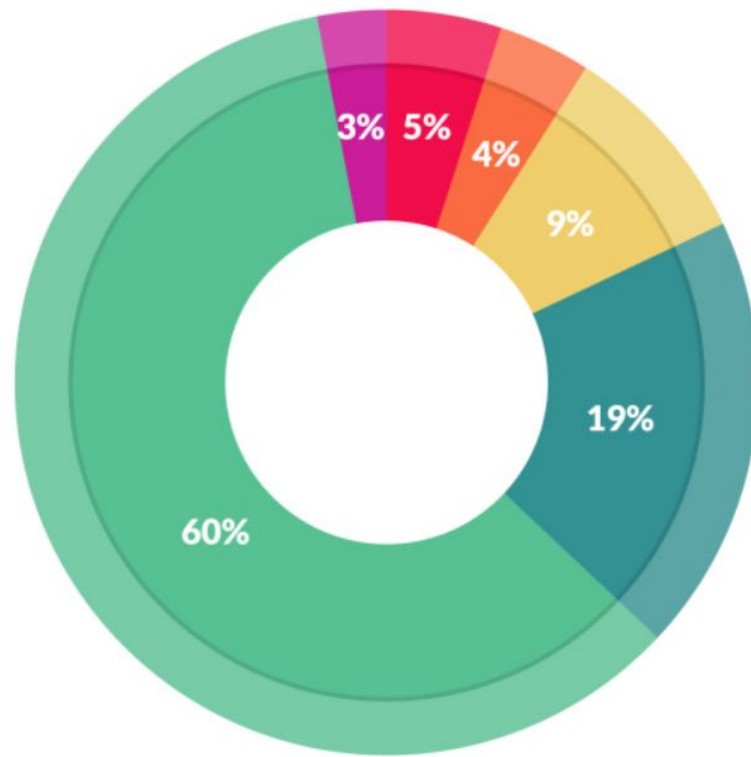


DSW

FAIR

DATA STEWARDSHIP WIZARD

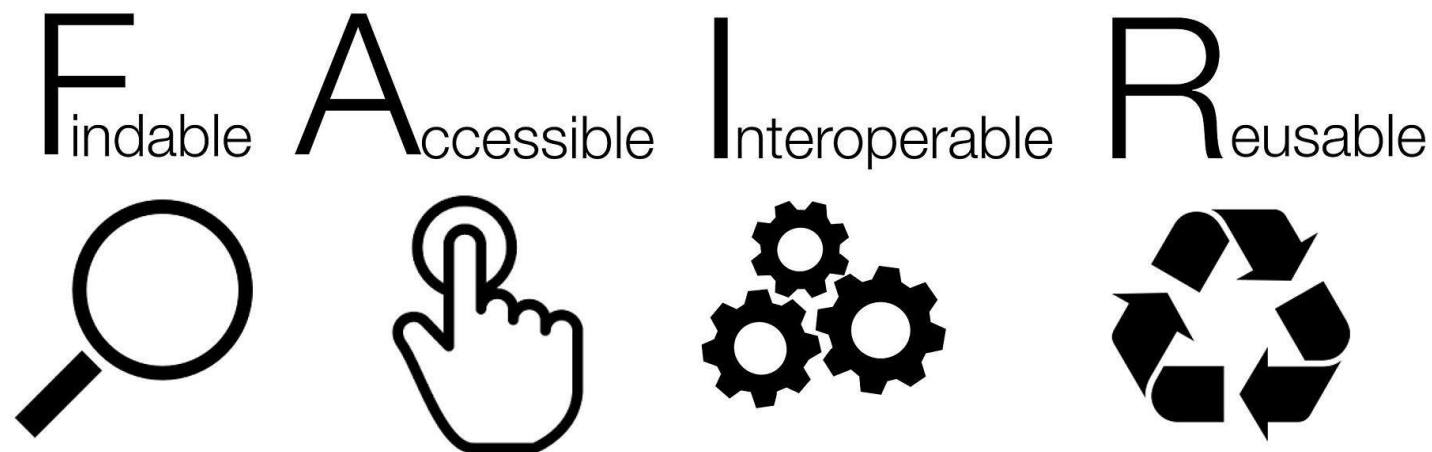




What data scientists spend the most time doing

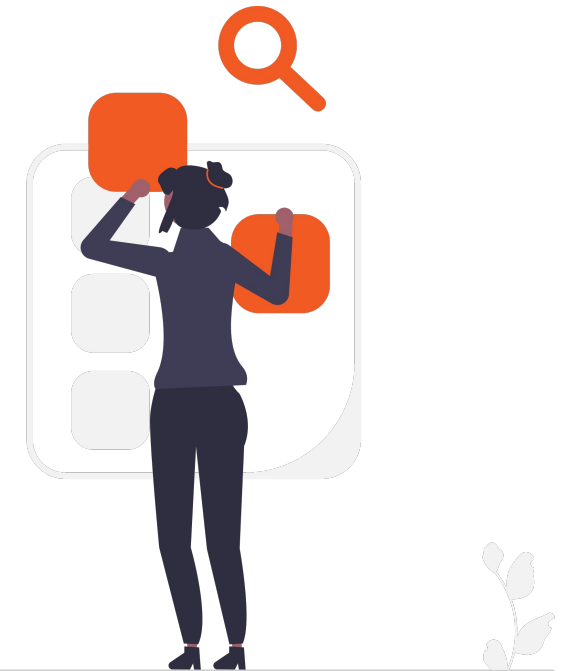
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

- Hlavním cílem je vylepšit **znovupoužitelnost vědeckých dat**
- Zdůrazňují strojovou zpracovatelnost (*machine-actionability*) a kvalitní metadata
- Formulováno [16 obecnými principy](#) (r. 2016)



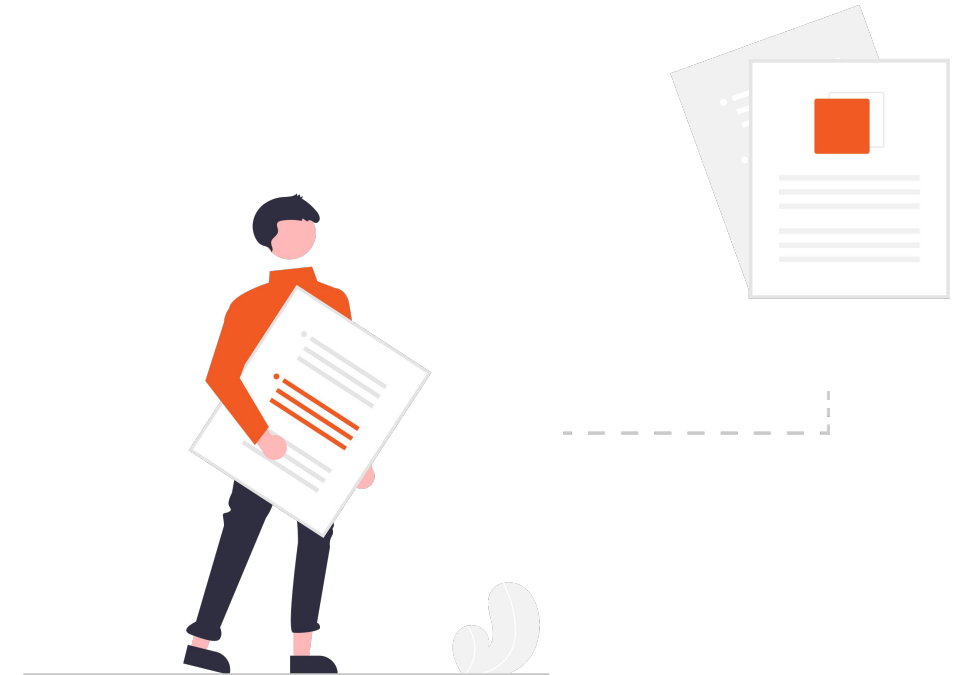
Findable (naležitelná)

- Prvním krokem ke znovupoužitelnosti je nalezení dat
- (Meta)data mají globálně unikátní persistentní identifikátor
- Data jsou popsána bohatými metadaty
- (Meta)data jsou registrována v prohledávatelném zdroji



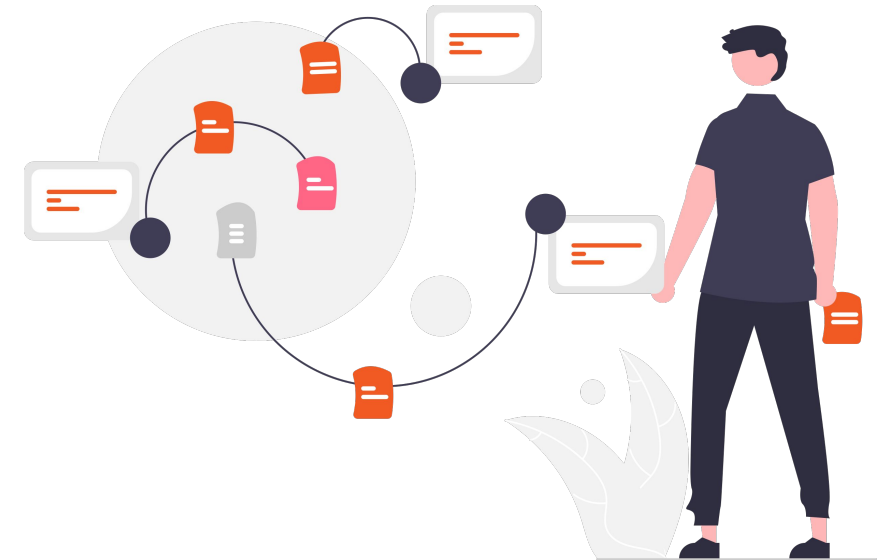
Accessible (dostupná)

- Po nalezení dat musí být jasné, jakým způsobem je k nim možné přistupovat
- **Accessible ≠ Open**
- Přístup k (meta)datům pomocí standardního komunikačního protokolu
 - Použití autentizace a autorizace pokud je potřeba



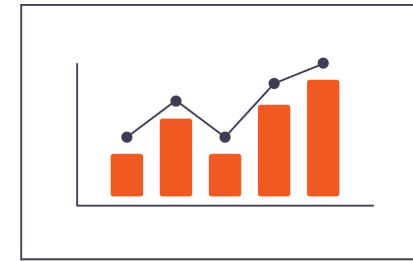
Interoperable (interoperabilní)

- Možnost integrace dat s jinými daty
- Možnost procesování v různých aplikacích a workflow
- Použití standardních formátů, slovníků a ontologií (RDF, JSON-LD, OWL)



Reusable (znovupoužitelná)

- Hlavním cílem FAIR je znovupoužitelnost
- Kvalitní popis (meta)dat
 - Licence
 - Původ (*provenance*)
 - Standardy komunity v dané doméně



- FAIR vize představuje "vrchol evoluce Homo Sapiens" v oblasti nakládání s daty
- Zaměřuje se na obecné technické a organizační aspekty
- Implementace FAIR principů je velké celosvětové téma se spoustou výzev – pionýrská doba se spoustou příležitostí



Kde se dozvědět více?



<https://www.go-fair.org>

DMP je běžná (a vyžadovaná) součást projektové žádosti



... a mnoho dalších

K čemu je dále dobré mít DMP



- Definuje role a zodpovědnosti pro práci s daty v týmu
- Pomůže při plánování prostředků a vybavení
- Pomůže identifikovat rizika a vybrat vhodná řešení na začátku projektu: "Prevence vs. hašení požárů"
- Usnadňuje sdílení, znovupoužitelnost a zachování dat (naplnění FAIR principů)
- I samotný proces plánování má svou hodnotu

"Plány jsou ničím; plánování je vším."



- Obecné informace o projektu
- Popis dat, která budou během projektu použita a vygenerována
- Použití metadat a ontologií, dokumentace
- Ukládání dat, bezpečnost a strategie pro zachování dat po ukončení projektu
- Sdílení dat
- Náklady a lidské zdroje potřebné pro správu dat
- Etické a právní otázky, licence
- Způsob naplnění FAIR principů

Jak vytvořit dobrý DMP?

Vytvořit dobrý DMP není snadné

Varianta 1: sám nastudovat vše potřebné

- Literatura
- Kurzy a tréninky

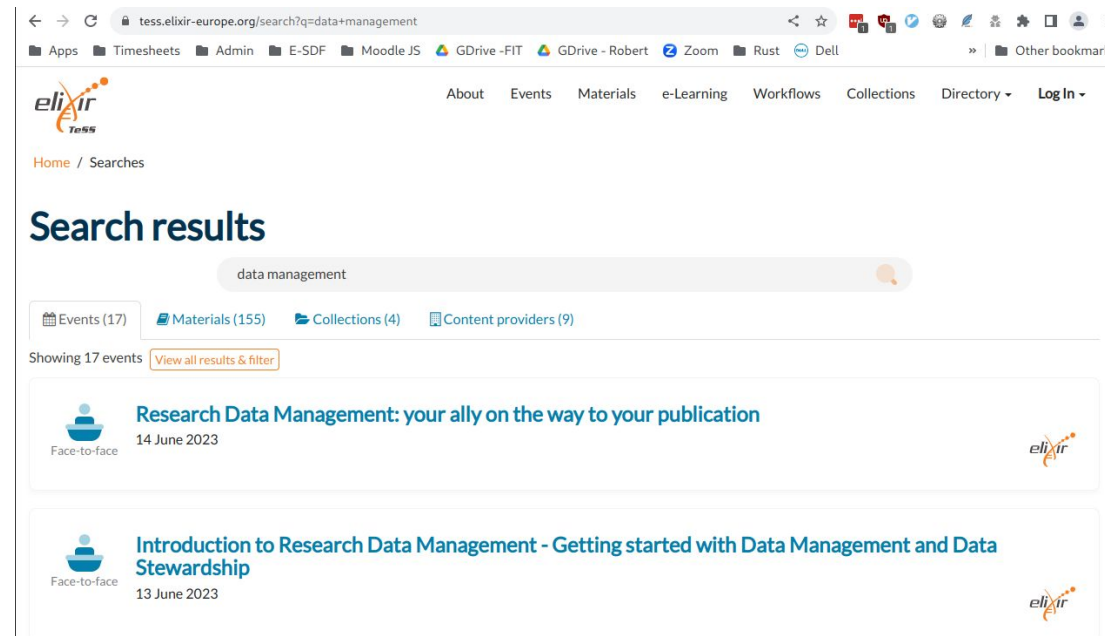


Book

Data Stewardship for Open Science

Implementing FAIR Principles
By *Barend Mons*

Edition	1st Edition
First Published	2018
eBook Published	27 February 2018
Pub. Location	New York
Imprint	Chapman and Hall/CRC
DOI	https://doi.org/10.1201/9781315380711
Pages	244
eBook ISBN	9781315380711
Subjects	Bioscience, Computer Science, Mathematics & Statistics



The screenshot shows a web browser window with the URL tess.elixir-europe.org/search?q=data+management. The page displays search results for 'data management' on the elixir Europe website. The search results are filtered to show 17 events. Two events are visible:

- Research Data Management: your ally on the way to your publication** (Face-to-face), dated 14 June 2023.
- Introduction to Research Data Management - Getting started with Data Management and Data Stewardship** (Face-to-face), dated 13 June 2023.

Jak vytvořit dobrý DMP?

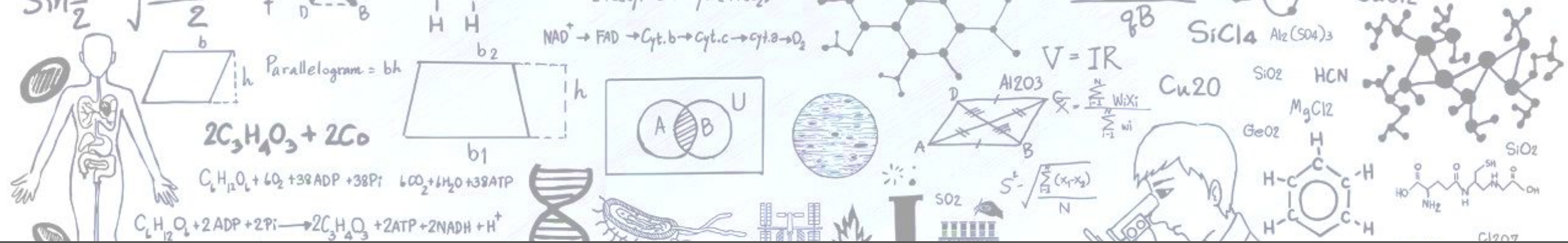


Varianta 2: vyhledat pomoc

A. Institucionální data steward

- Ne každá instituce zatím má
- Bývá vytížen

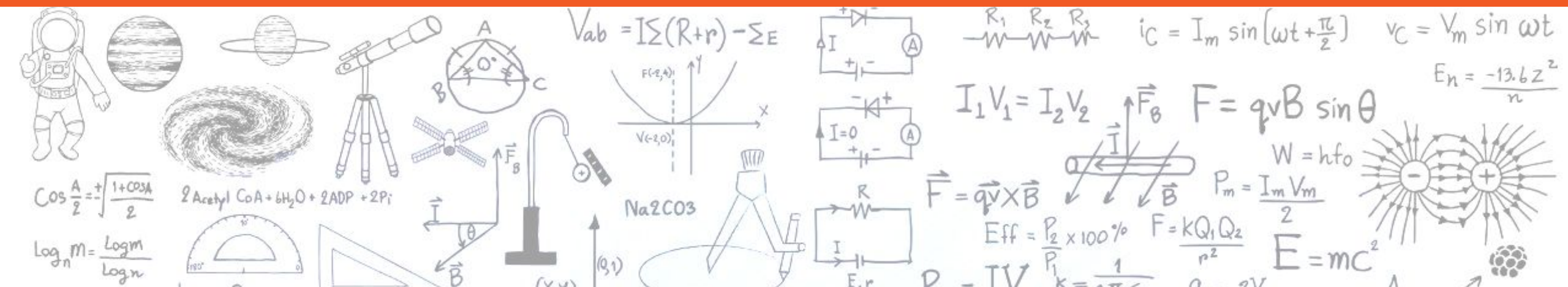
B. Data Stewardship Wizard = elektronický data steward se spoustou možností a funkcí, který je k dispozici 24x7



DSW

Data Stewardship Wizard

DATA STEWARDSHIP WIZARD



- Open-source vyvíjený ve spolupráci **ELIXIR CZ** a **ELIXIR NL**
- Expertní systém pro plánování správy dat a tvorbu data management plánů
 - *“From burden to benefit”*
- Doporučený nástroj mj. v **Horizon Europe Program Guide**.

Hlavní myšlenky DSW



- Minimum psaní = plán není eseje, psaní jen tam, kde je to nezbytné
- Vedení = DSW vede uživatele skrz tzv. *smart questionnaire*
- Flexibilita = lze upravovat obsah a integrovat s jinými službami
- Otevřenost = kdokoli je může využít a vytvářet vlastní obsah
- Orientace na uživatele = vývoj se výrazně řídí na základě zpětné vazby

DMP jako předletová příprava



Operations Checklist		
Parking Brake		
Fuel Flow	Set	
Battery Switch	Cutoff	
Hydraulic Pump	ON	
Landing Gear	On	
Flaps	Check	✓
Spoiler	Up	✓
Fuel Amount	Retracted	✓
De-Ice	Check	✓
Passenger Sign	Off	✓
Check Weather	Off	✓
	Flight Services	
Transponder		
Anti Collision Ligths	Standby	
	On	
Engine Start Switches		
Thrust Reverser Switch	Check	
Master Switch	On	

Dotazník (Smart Questionnaire)



- Interaktivní formulář pro získávání všech informací důležitých pro DMP
- Odpovídat je možné v libovolném pořadí
- Jsou zobrazovány pouze relevantní otázky na základě předchozích odpovědí

The screenshot displays the 'Smart Questionnaire' interface. On the left, a sidebar shows the 'Current Phase' as 'Before Submitting the Proposal' and a list of 'Chapters' with their respective question counts:

Chapter	Count
I. Administrative information	4
II. Re-using data	4
III. Creating and collecting data	10
IV. Processing data	8
V. Interpreting data	9
VI. Preserving data	10
VII. Giving access to data	2

The main content area shows a question titled '1.b.1 Will you be using any pre-existing data (including other people's data)?'. The question text asks: 'Will you be referring to any earlier measured data, reference data, or data that should be mined from existing literature? Your own data as well as data from others?'. Below the question, there are two radio button options: 'a. No' and 'b. Yes', with 'b. Yes' selected. A 'Clear answer' button is also present. Below this question, another question titled '1.b.1.b.1 What reference data will you use?' is visible, with its text: 'Much of todays data is used in comparison with reference data. You may be comparing your own data with a "standard set" which is maintained as a collection by someone else. Or you could be determining differences to a standard (for example in bioinformatics, a genome is often compared with a reference genome to identify genomic variants). If you use reference data, there are several specific issues that you should consider. What are the reference data sets that you will use?'. A 'Data Stewardship for Open Science' link is provided for each question, with 'ezi' for the first and 'quc' for the second. An '+ Add' button is located at the bottom of the question list.

Název → **1.a.4.b.1.a.1 What repository will this data be stored in?**

Popis → Domain repositories often have the best functionality to make the data findable and reusable: even though it may look like a database that could be reused in a completely different field would be better findable in a generic repository, the limited availability of domain-specific metadata make that less valuable.

Many repositories are listed in <https://fairsharing.org/>

If a repository offers to give your data set a DOI or alternative persistent identifier it is a good idea to use that option.

External links: [FAIRSharing](#), [Registry of Research data Repositories](#) → **Reference**

Možné odpovědi →

- a. A domain-specific repository **Findability**
- b. Our national repository **Findability** → **FAIR metriky**
- c. Our institutional repository **Findability**
- d. A special-purpose repository for the project **Findability**

Kdo odpověděl → Answered in less than 5 seconds by Albert Einstein.

Informace k odpovědi → Disadvantage of a general purpose repository is the lack of data-specific features (e.g. 'play' instead of 'download' for an audio file) and limited findability

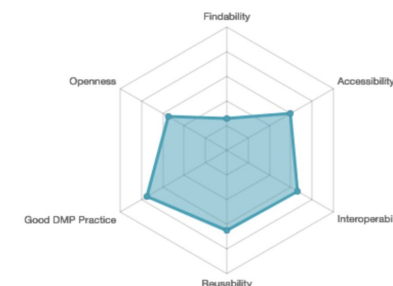
- Nejsou vysloveně "dobré a špatné" odpovědi, některé volby ale mohou být vhodnější než jiné.
- DSW poskytuje zpětnou vazbu formou metrik - F, A, I, R, G, O
- Výsledná hodnota se počítá jako vážený průměr všech odpovědí, které danou metriku ovlivňují

Summary Report

Answered (current phase): 15/38

Answered: 41/101

Metric	Measure
Findability	0.26
Accessibility	0.60
Interoperability	0.67
Reusability	0.65
Good DMP Practice	0.75
Openness	0.55



Online spolupráce



Albert Einstein Owner ✕

Nikola Tesla Editor ✕

Isaac Newton Viewer ✕

Visible by all other logged-in users
Other logged-in users can **edit** the Project.

Public link
Anyone with the link can **view** the Project.

<https://demo.ds-wizard.org/projects/7aaec?> **Copy link**

ew Documents Settings

Comments 7 TODOs 1 Version history

1.a.1 Data type + 1 comment

Desirable: Before Submitting the Proposal

XML

Answered in less than 5 seconds by Albert Einstein.

1.a.2 How is this data structured?

Desirable: Before Submitting the Proposal

- a. A structured domain specific file with data and metadata fields
- b. A table or set of tables (consisting of 'data records')
- c. Complex data, like a graph

View resolved comments

Comments 1 Editor notes

Anonymous user
1. 10. 2021, 14:49

Do we really want to use XML?

Reply...

Create a new comment...

Vygenerování dokumentu s DMP



K dispozici je řada šablon na "push of button"

- Horizon 2020 DMP
- Horizon Europe DMP
- Science Europe DMP
- Machine-actionable DMP (RDA Common Standard)
- Questionnaire Report

Data Management Plan

Science Europe Example

Contact person: [Jana Freeman \(jana.freeman@ds-wizc.cz\)](mailto:jana.freeman@ds-wizc.cz)
0000-0000-0000-0001
[Czech Technical University in Prague](http://www.ctu.cz)

Based on: Common DSW Knowledge Model, 2.3.0

Created by: [Jana Freeman \(jana.freeman@ds-wizc.cz\)](mailto:jana.freeman@ds-wizc.cz)
DSW

Generated on: 24 Jun 2021

Projects

We will be working on the following projects and for those described in this DMP.

Arsenic and Selenium Speciation Using Hyphenated

Start date: 1.1.2021
End date: 31.12.2021
Funding: [Grantová Agentura České Republiky: grant \(planned\)](#)

The main goal of this study is to determine whether arsenic and selenium are produced from the meadows in the vicinity of old metal mines in Příbram, Kutná Hora, and Nalžovské Hory (Czech Republic) by herbivorous herds. Total and speciation analysis of As and Se will be conducted using a hyphenated technique of HPLC and ICP-MS.

Section A: Data Collection

1. What data will you collect or create?

Instrument datasets

The following instrument datasets will be acquired in the project:

- **HPLC**
This dataset will be collected by experts in the project using HPLC equipment.
The equipment is very well described and known.
- **ICP-MS**
This dataset will be collected by experts in the project using ICP-MS equipment.
The equipment is very well described and known.

Re-used datasets

We will use the following reference datasets:

- [Chemical Component Dictionary \(http://dx.doi.org/10.26434/chemrxiv-2019-08-01\)](http://dx.doi.org/10.26434/chemrxiv-2019-08-01)

We will use the following already existing non-reference datasets:

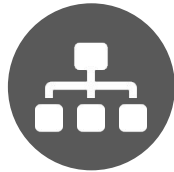
- **Previous in-house Arsenic and Selenium analysis**
We already have a copy of this dataset.

Data formats and types

We will be using the following data formats and types:

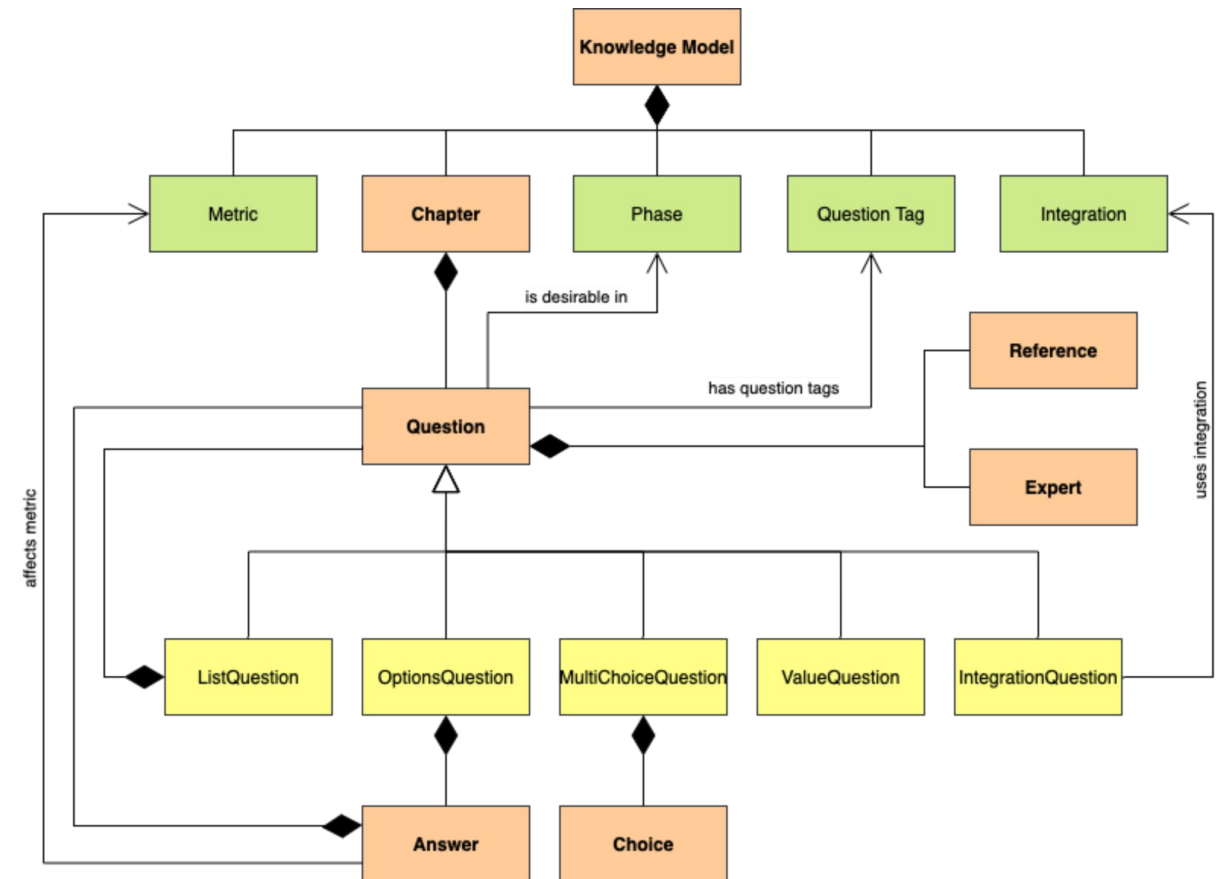
- [Chemistry vocabulary](#)
It is a standardized format. This is a suitable format for data storage. We will have only a small amount of data stored in this format.

Knowledge Model



Knowledge Model (KM)

- Obsahuje **znalosti** o tom, na co a jak je potřeba se ptát
- **Šablona** pro strukturovaný dotazník (questionnaire)
- Stromová struktura sestávající z **kapitol, otázek, odpovědí**, navazujících otázek a dalších zdrojů



Knowledge Model Editor



Common DSW Knowledge Model 🔍

Publish

Knowledge Model Question Tags Preview Settings

... > What existing reference data... > Will you be using this referen... > Yes > What are the conditions of u...

Expand all Collapse all

- Common DSW Knowledge Model
 - Administrative information
 - Re-using data
 - Is there any pre-existing data?
 - No
 - Yes
 - What existing reference data did you consider re-using?
 - Reference database or dataset
 - Where is this reference data available
 - Will you be using this reference data set?
 - No
 - Yes
 - Name and contact details of the owner of this data
 - What are the conditions of use for this database or dataset?
 - They are freely available for any use (public domain or CC0)
 - They are freely available with obligation to quote the source (e.g. Wikipedia on Copyright)
 - They are available under some restrictions, which we will follow in
 - Do you know in what format the reference data is available?

Question

aee03da5

Move

Delete

Type

Options

Title

What are the conditions of use for this database or dataset?

Text

Editor

Preview

Although there is no world-wide rule for the application of copyright on data sets (copyright only applies to things that require a so-called "creative step"), it is wise to check for an explicit permission to use a data

- Hodnota (Value) – text, email, datum, číslo,...
- Výběr odpovědi (Options) – umožňuje větvení dotazníku
- Výběr z více možností (Multi-Choice)
- Seznam položek (List of items) – každá položka má stejné odpovědi
- Integrace (Integration) – propojení s externí službou

Možnosti přizpůsobení KM

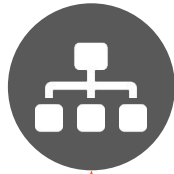


- Rozšíření nebo úpravy existujících KM
- Vhodné pro rozšíření pro jiné domény nebo instituce
- Příklad
 - **Common DSW Knowledge Model** – obecný KM pro data stewardship
 - **Life Sciences DSW Knowledge Model** – rozšiřuje obecný KM o otázky z Life Sciences domény

Data Stewardship Wizard



Knowledge Model

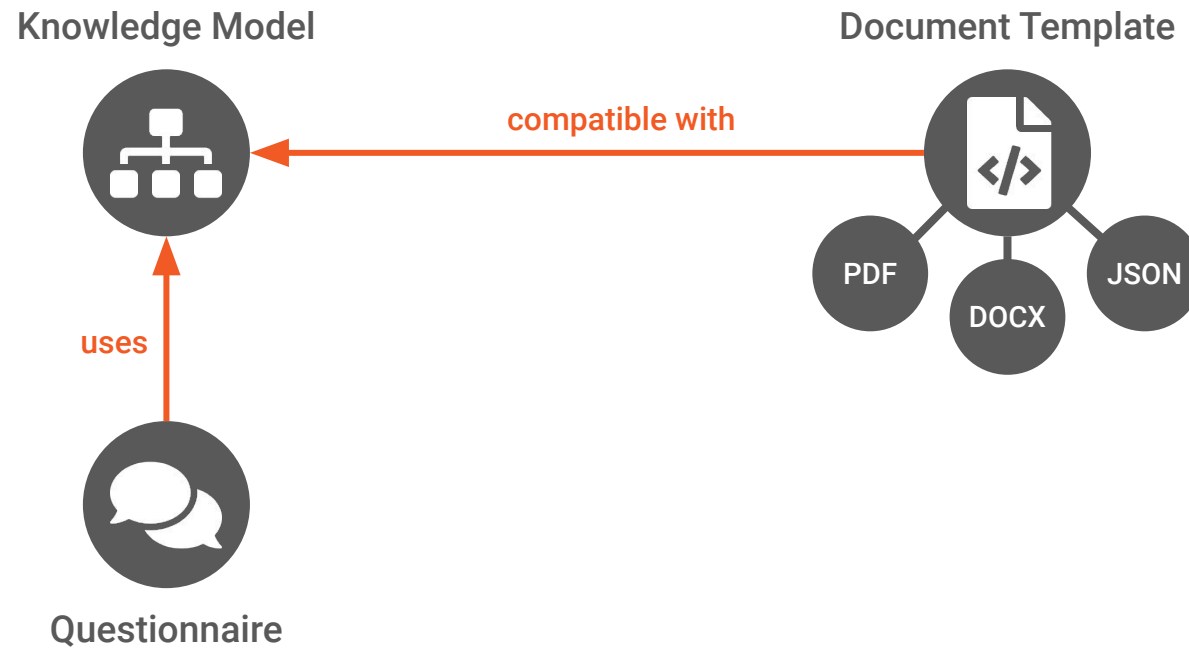


uses



Questionnaire

Data Stewardship Wizard



Šablona pro dokumenty (Document Template)



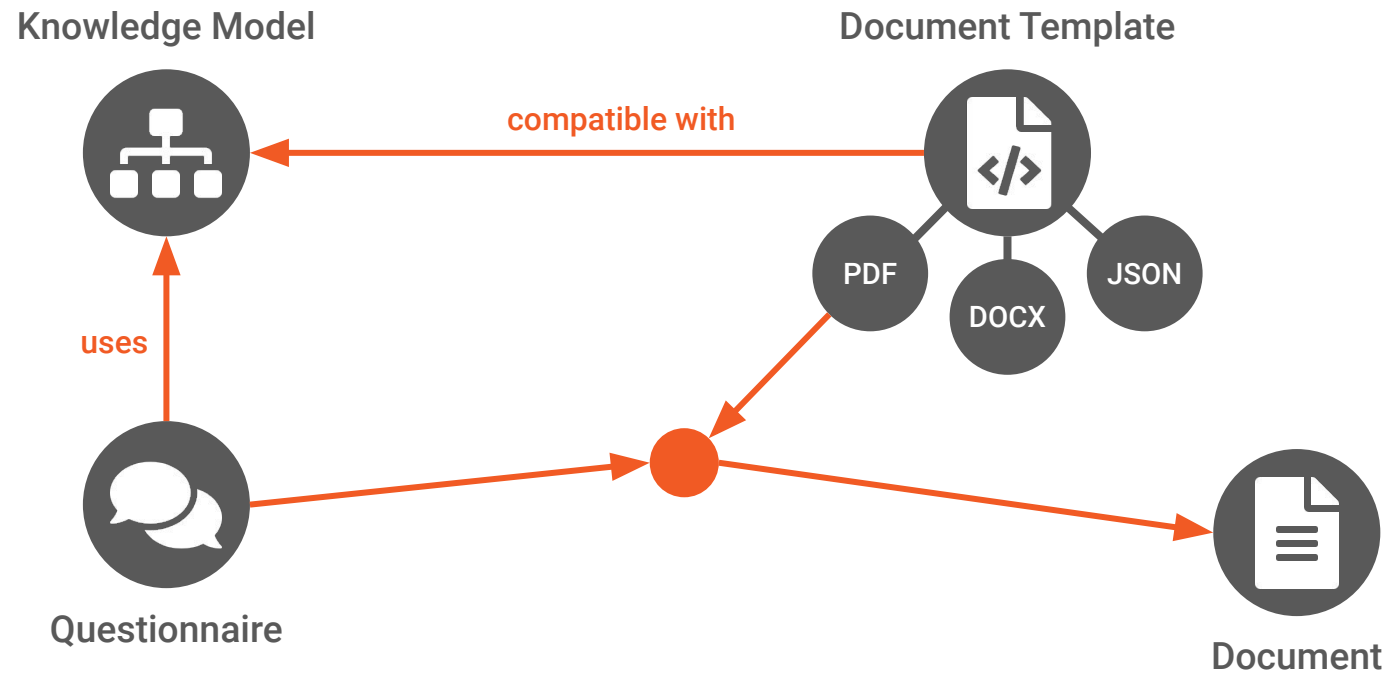
- Tvoření šablon vyžaduje více **technických dovedností**
- Šablony jsou složeny z **JSON** metadat, **Jinja2** šablon a dalších souborů
- Tvoření usnadňuje **DSW Template Development Kit (TDK)**

A screenshot of a code editor showing a file named 'template.json'. The left sidebar shows a file explorer with a folder 'MADMP-TEMPLATE-MA...' containing files like '_mapping.j2', '_uuids.j2', 'madmp.json.j2', 'madmp.ttl.j2', '.gitignore', 'LICENSE', 'README.md', and 'template.json'. The main editor area shows the JSON content of 'template.json' with line numbers 1 through 16. The JSON is a metadata object for a document template.

```
1 {  
2   "organizationId": "dsw",  
3   "templateId": "rda-madmp",  
4   "version": "1.4.0",  
5   "name": "maDMP (RDA DMP Comm",  
6   "description": "Machine-acti",  
7   "recommendedPackageId": "dsw",  
8   "license": "Apache-2.0",  
9   "metamodelVersion": 3,  
10  "allowedPackages": [  
11    {  
12      "orgId": "dsw",  
13      "kmId": "root",  
14      "minVersion": "2.3.0",  
15      "maxVersion": null  
16    },  
17  ]  
18 }
```



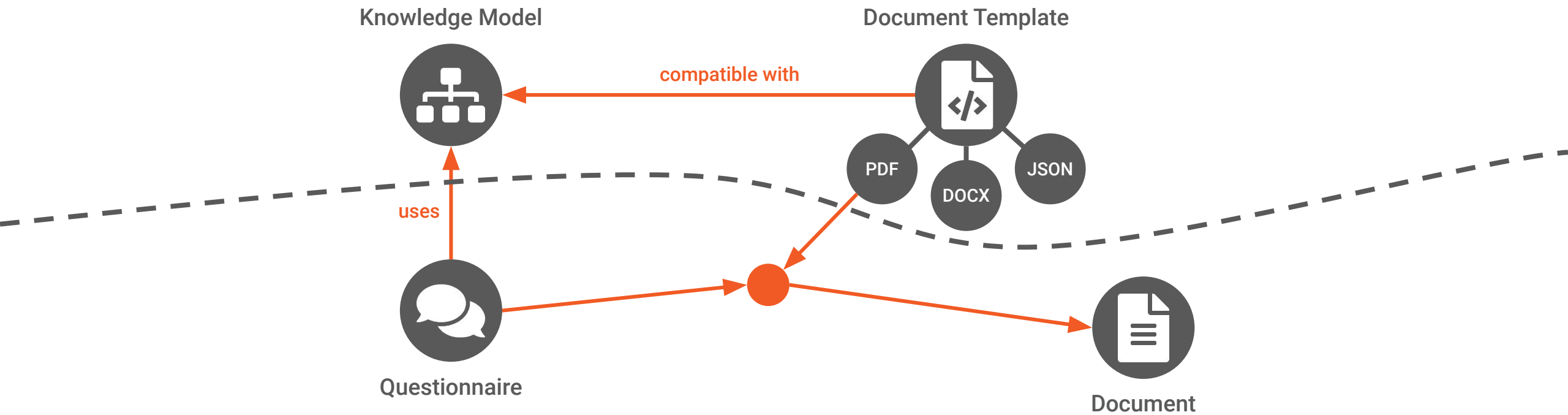
Data Stewardship Wizard



Data Stewardship Wizard



Data Steward



Researcher

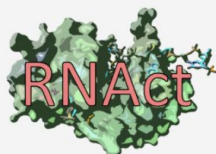
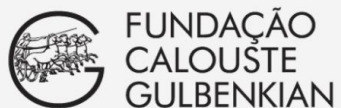
A mnohem víc



- Data stewardship hints
- Project phases
- Project templates
- Editor notes
- Version history
- Migrace knowledge modelů & aktualizace dotazníků
- Integration widget
- Submission service
- Single-Sign On
- ...

Kde se DSW používá

BioData.pt



DSW a další na Masarykově univerzitě



- DSW MUNI – univerzitní instance DSW: <https://dsw.muni.cz/>
- Open Science web MU: <https://openscience.muni.cz/>
 - [Návod na použití DSW MU s příklady DMP z různých oborů.](#)
 - Kontakty na Open Science core tým a fakultní metodiky: <https://openscience.muni.cz/kontakty>
- Vzdělávací aktivity:
 - Kurz [Bezpečná správa výzkumných dat](#) (chystáme třetí běh kurzu; datum bude upřesněno).
 - Primárně pro zájemce z řad výzkumných pracovníků, projektové a vědecké podpory.
 - Kurz [CORE042: Data – odpověď na základní otázku života, vesmíru a vůbec...](#)
 - Primárně určen pro bakalářské studenty všech fakult MU.
 - Vytvořeno jako zcela nový předmět v rámci vznikajících kurzů [Společného univerzitního základu](#) pro rozšiřování znalostí studentů za hranice jejich domovského oboru.
 - Chceme zasévat semínka pro výchovu nové generace Data Stewards.

DS Wizard | Masaryk University Log In

Log In

Email

Password

[Forgot your password?](#) Log In

Or connect with

↔ Jednotné přihlášení MUNI

Poděkování



Provoz a vývoj nástroje DSW je podporován infrastrukturou ELIXIR CZ
(MŠMT grant č.: LM2023055).

Děkuji za pozornost



- **Otázky a diskuse**

- Data Stewardship Wizard

<https://ds-wizard.org>

 [@dswizard_org](https://twitter.com/dswizard_org)

- Robert Pergl

robert.pergl@ds-wizard.org

